AP16-OL7: A Multilingual Database for Oriental Languages and A Language Recognition Baseline

Dong Wang*, Lantian Li*, Difei Tang† and Qing Chen†
* CSLT, Tsinghua University
E-mail: {wangdong99,lilt13}@mails.tsinghua.edu.cn
† SpeechOcean
E-mail:{tangdifei,chenqing}@speechocean.com

***Abstract*—We present the AP16-OL7 database which was released as the training and test data for the oriental language recognition (OLR) challenge on APSIPA 2016. Based on the database, a baseline system was constructed on the basis of the i-vector model. We report the baseline results evaluated in various metrics defined by the AP16-OLR evaluation plan and demonstrate that AP16-OL7 is a reasonable data resource for multilingual research.**

I. Introduction

Oriental languages, including various languages spoken in east, northeast and southeast Asia, belong to several language families, including Austroasiatic languages (e.g.,Vietnamese, Cambodia) [1], TaiKadai languages (e.g., Thai, Lao), Hmong-Mien languages (e.g., some dialects in south China), Sino-Tibetan languages (e.g., Chinese Mandarin), Altaic languages (e.g., Korea, Japanese), Indo-European languages (e.g., Russian) [2], [3], [4]. These languages are generally believed to be genetically unrelated and were developed from diverse cultures. However, they do share many features due to the demographic migration and international business interaction in history. For example, many languages in the so-called Mainland Southeast Asia (MSEA) linguistic area posses a particular syllable structure that involves monosyllabic morphemes, lexical tone, a fairly large inventory of consonants [5]. Another example is the significant influence of Chinese to Korean, Japanese, Vietnamese and many languages in southeast Asia. In the modern period, English becomes the most influential language, resulting in numerous English-originated words in almost all oriental languages.

The complex acoustic and linguistic patterns of oriental languages have attracted much interest in a multitude of research areas, including comparative phonetics, evolutionary linguistics, second language acquisition, and social linguistics. In particular, the diverse evolution paths of these languages and their complicated interaction offers a valuable opportunity for studying mixlingual and multilingual phenomena.

Despite the broad interest, data resources of oriental languages are far from abundant. One possible reason is that many of these languages are spoken by a relatively small population, and most of the speakers are in developing countries. Some effort has been devoted to building data resources for oriental languages, e.g., the annual oriental COCOSDA (OC) workshop intends to promote speech and language resource construction for oriental languages, and the transactions on Asian and Low-Resource Language Information Processing (TALLIP) journal calls for original research on oriental languages, especially languages with limited resources.[1] Some projects, e.g., the Babel program[2], although not particularly for oriental languages, do involve Vietnamese, Thais, Lao and some other low-resource languages in southeast Asia. In spite of these efforts, resource construction and corresponding research on oriental languages are still rather limited, except one or two rich-resource languages, such as Chinese and Japanese.

To promote research for oriental languages, particularly on multilingual speech and language processing, the center for speech and language technologies (CSLT) at Tsinghua University and Speechocean collaborated together and organized an oriental language recognition (OLR) challenge on APSIPA 2016. This event called for a competition on a language recognition task on seven oriental languages. To support this event, Speechocean released a multilingual speech database AP16-OL7 and made it free for the challenge participants. This paper will present the data profile of the database, the evaluation rules of the challenge, and a baseline system that the participants can refer to.

Note that there are several databases that can be used for multilingual research. For example, polyphone [6], globalPhone [7], NTT multilingual database[3], SPEECHDAT-CAR [8],Speechdat-E [9], Babel [10], and the multilingual databases created by the new Babel project. To our best knowledge, AP16-OL7 is the first multilingual speech database specifically designed for oriental languages.

II. Database profile

The AP16-OL7 database was originally created by Speechocean targeting for various speech processing tasks (mainly speech recognition). The entire database involves seven datasets, each in a particular language. The seven languages are: Mandarin, Cantonese, Indoesian, Japanese, Russian, Korean, Vietnamese. The data volume for each language is about 10 hours of speech signals recorded by 24 speakers (12 males

[1] https://mc.manuscriptcentral.com/tallip
[2] https://www.iarpa.gov/index.php/research-programs/babel
[3] http://www.ntt-at.com/product/speech2002/

TABLE I
AP16-OL7 Data Profile

Datasets			Training & Dev set			Test set		
Code	Description	Channel	No. of Speakers	Utt./Spk.	Total Utt.	No. of Speakers	Utt./Spk.	Total Utt.
ct-cn	Cantonese in China Mainland and Hongkong	Mobile	18	320	5759	6	320	1920
zh-cn	Mandarin in China	Mobile	18	300	5398	6	300	1800
id-id	Indonesian in Indonesia	Mobile	18	320	5751	6	320	1920
ja-jp	Japanese in Japan	Mobile	18	320	5742	6	320	1920
ru-ru	Russian in Russia	Mobile	18	300	5390	6	300	1800
ko-kr	Korean in Korea	Mobile	18	300	5396	6	300	1800
vi-vn	Vietnamese in Vietnam	Mobile	18	300	5400	6	300	1800

Male and Female speakers are balanced.
The number of total utterances might be slightly smaller than expected, due to the quality check.

and 12 females), and each speaker recorded about 300 utterances in reading style. The signals were recorded by mobile phones, with a sampling rate of 16kHz and a sample size of 16 bits. Each dataset was split into a training set consisting of 18 speakers, and a test set consisting of 6 speakers. For Mandarin, Cantonese, Vietnamese, Russian and Indonesia, the recording was conducted in a quiet environment. As for Korean and Japanese, there are 2 recording sessions for each speaker: the first session was recorded in a quiet environment and the second was recorded in a noisy environment. The basic information of the AP16-OL7 database is presented in Table I.

Besides the speech signals, the AP16-OL7 database also provides lexicons of all the seven languages, and transcriptions of all the training utterances. These resources allow training acoustic-based or phonetic-based language recognition systems. Training phone-based speech recognition systems is also possible, though large vocabulary recognition systems are not well supported, due to the lack of large-scale language models.

The AP16-OL7 database is freely available for the participants of the AP16-OLR challenge and the APSIPA 2016 special session on *multilingual speech and language processing*. It is also available for any academic and industrial users, subject to a slightly different licence from SpeechOcean.[4]

III. AP16-OLR challenge

Based on the AP16-OL7 database, we call an oriental language recognition (OLR) challenge.[5] Following the definition of NIST LRE15 [11], the task of the challenge is defined as follows: Given a segment of speech and a language hypothesis (i.e., a target language of interest to be detected), the task is to decide whether that target language was in fact spoken in the given segment (yes or no), based on an automated an analysis of the data contained in the segment. The AP16-OLR evaluation plan also follows the principles of NIST LRE15: it focuses on the close-set condition, and allows no additional training materials besides AP16-OL7. The evaluation details are described as follows.

A. System input/output

The input to the OLR system is a set of speech segments in unknown languages (but within the 7 languages of AP16-OL7). The task of the OLR system is to determine the confidence that a language is contained in a speech segment. More specifically, for each speech segment, the OLR system outputs a score vector $< \ell_1, \ell_2, ..., \ell_7 >$, where ℓ_i represents the confidence that language i is spoken in the speech segment. Each score ℓ_i will be interpreted as follows: if $\ell_i \geq 0$, then the decision would be that language i is contained in the segment, otherwise it is not. The scores should be comparable across languages and segments. This is consistent with the principle of LRE15, but differs from that of LRE09 [12] where an explicit decision is required for each trial.

In summary, the output of an OLR submission will be a text file, where each line contains a speech segment plus a score vector for this segment, e.g.,

seg_1	0.5	-0.2	-0.3	0.1	-9.2	-0.1	-5.1
seg_2	-0.1	-0.3	0.5	0.3	-0.5	-0.9	-3.2
...				...			

B. Test condition

- No additional training materials are allowed to use.
- All the trials should be processed. Scores of lost trials will be interpreted as -inf.
- Each test segment should be processed independently. Knowledge from other test segments is not allowed to use (e.g., score distribution of all the test segments).
- Information of speakers is not allowed to use.
- Listening to any speech segments is not allowed.

C. Evaluation metrics

As in LRE15, the AP16-OLR challenge chooses C_{avg} as the principle evaluation metric. First define the pair-wise loss that composes the missing and false alarm probabilities for a particular target/non-target language pair:

$$C(L_t, L_n) = P_{Target} P_{Miss}(L_t) + (1 - P_{Target}) P_{FA}(L_t, L_n)$$

where L_t and L_n are the target and non-target languages, respectively; P_{Miss} and P_{FA} are the missing and false alarm probabilities, respectively. P_{target} is the prior probability for the target language, which is set to 0.5 in the evaluation. Then the principle metric C_{avg} is defined as the average of the above pair-wise performance:

[4] http://speechocean.com

[5] http://cslt.riit.tsinghua.edu.cn/mediawiki/index.php/ASR-events-AP16-details

$$C_{avg} = \frac{1}{N} \sum_{L_t} \left\{ \begin{array}{l} P_{Target} \cdot P_{Miss}(L_t) \\ + \sum_{L_n} P_{Non-Target} \cdot P_{FA}(L_t, L_n) \end{array} \right\}$$

where N is the number of languages, and $P_{Non-Target} = (1 - P_{Target})/(N - 1)$.

IV. Baseline results

We present baseline language recognition systems based on the i-vector model, and evaluate the performance in terms of the metrics defined by the AP16-OLR challenge. The purpose of these experiments is not to present a competitive submission, instead to demonstrate that the AP16-OL7 database is a reasonable data resource to conduct language recognition research.

A. Database

The AP16-OL7 test data was not released when this paper was written, we therefore randomly selected a subset of the training data as the training set to develop the system, and the remaining as the test set to evaluate system performance. The splitting was conducted according to speakers. After the splitting, the training set involved 15 speakers, and the test set involved 3 speakers. In order to evaluate potential performance variation caused by data splitting, the splitting was conducted three times, giving rise to three data configurations each consisting of different of training and test data, denoted by '$G1$', '$G2$', '$G3$', respectively.

B. Experimental setup

The baseline system was constructed based on the i-vector model [13], [14]. The static acoustic features involved 19-dimensional Mel frequency cepstral coefficients (MFCCs) and the log energy. This static features were augmented by their first and second order derivatives, resulting in 60-dimensional feature vectors. The UBM involved $2,048$ Gaussian components and the dimensionality of the i-vectors was 400. Linear discriminative analysis (LDA) was employed to promote language-related information. The dimensionality of the LDA projection space was set to 6.

With the i-vectors (either original or after LDA transform), the score of a trail on a particular language can be simply computed as the cosine distance between the test i-vector and the mean i-vector of the training segments that belong to that language. This is denoted to be 'cosine distance scoring'. A more powerful scoring approach is to employ various discriminative models. In our experiment, we trained a support vector machine (SVM) for each language to determine the score that a test i-vector belongs to that language. The SVMs were trained on the i-vectors of all the training segments, following the one-verse-rest scheme. We will call this scoring approach as 'SVM-based scoring'.

C. Visualization with T-SNE [15]

To provide an intuitive understanding of the discriminative capability of i-vectors on languages, the i-vectors of all the segments in the test set are plotted in a two-dimensional space via T-SNE [15]. To save space, we present the results with the data configuration $G1$ only. Fig. 1 shows the original i-vectors, and Fig. 2 shows the i-vectors after LDA transform, where each color/shap represents a particular language. It can be seen that for the original i-vectors, each language is split into several clusters basically due to different speakers. After LDA, speaker information is suppressed and the language identify is more significant.

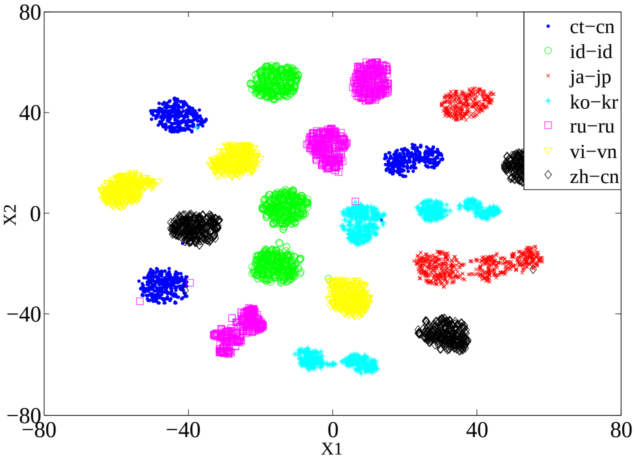

Fig. 1. Original i-vectors plotted by t-SNE. Each color/shape represents a particular language.

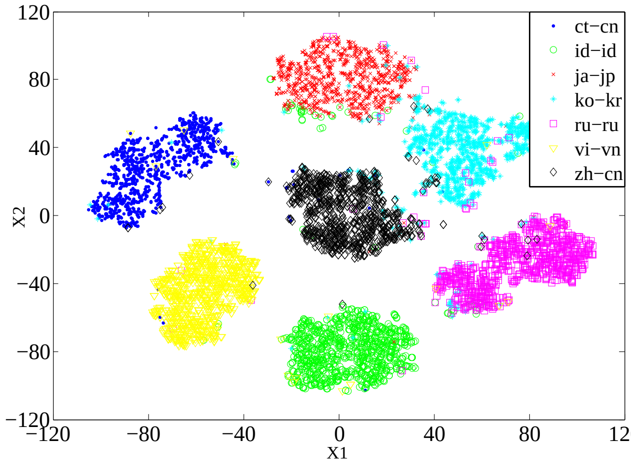

Fig. 2. LDA-transformed i-vectors plotted by t-SNE. Each color/shape represents a particular langauge.

D. Performance results

The primary evaluation metric in AP16-OLR is C_{avg}. Besides that, we also present the performance in terms of equal error rate (EER), minimum detection cost function (minDCF), detection error tradeoff (DET) curve, and identification rate (IDR). These metrics evaluate the system from different perspectives, offering a whole picture of the verification/identification capability of the baseline system.

1) C_{avg} results: The C_{avg} results for the three data splitting configurations and their average are shown in Table II. The rows 'i-vector' and 'LDA-vector' present the results with the cosine distance scoring; 'i-vector-SVM' and 'LDA-vecotr-SVM' present the results with the SVM-based scoring. 'Linear', 'Poly', and 'RBF' represent the three commonly used kernel functions. It can be seen that LDA leads to consistent performance gains, and the SVM-based scoring tends to outperform cosine distance scoring. Another observation is that the results with the three data profiles ($G1$/$G2$/$G3$) vary in a significant way, but the relative performance of different methods keeps stable.

TABLE II
C_{avg} RESULTS OF VARIOUS BASELINE SYSTEMS

	C_{avg} * 100			
System	$G1$	$G2$	$G3$	AVE
i-vector	4.97	7.20	6.65	6.27
LDA-vector	3.48	5.44	5.24	4.72
i-vector-SVM (Linear)	4.71	6.93	5.98	5.87
i-vector-SVM (Poly)	3.72	6.75	5.43	5.30
i-vector-SVM (RBF)	3.30	5.94	4.55	4.60
LDA-vector-SVM(Linear) (Linear)	3.16	5.40	5.27	4.61
LDA-vector-SVM(Poly) (Poly)	3.48	5.58	5.33	4.80
LDA-vector-SVM(RBF) (RBF)	3.31	5.53	5.25	4.70

2) EER and minDCF results: Besides C_{avg}, EER and minDCF are also widely used in measuring performance of verification systems. Compared to C_{avg}, these two metrics are not related to the decision result, but the quality of the scoring, and therefore evaluate the verification system from a different angle. The results for these two metrics are presented in Table III and Table IV respectively. It can be seen that similar conclusions can be drawn from these results as from the C_{avg} results.

E. DET curve

The DET curve is another popular way to evaluate verification systems. Compared to C_{avg}, EER and minDCF, the DET curve presents performance on all operation points, and therefore can evaluate a verification system in a more systematic way. To make the presentation clear, only the results based on the data profile $G1$ are presented, as shown in Fig 3. The black circles represent the operation location where the minDCFs are obtained. Again, similar conclusions as with the C_{avg}, EER and minDCF can be obtained.

TABLE III
EER RESULTS OF VARIOUS BASELINE SYSTEMS

	EER%			
System	$G1$	$G2$	$G3$	AVE
i-vector	5.70	8.21	7.97	7.29
LDA-vector	4.06	6.08	5.66	5.27
i-vector-SVM (Linear)	4.71	7.10	5.94	5.92
i-vector-SVM (Poly)	3.70	6.89	5.40	5.33
i-vector-SVM (RBF)	3.30	6.05	4.55	4.63
LDA-vector-SVM (Linear)	3.16	5.50	5.28	4.65
LDA-vector-SVM (Poly)	3.47	5.66	5.26	4.80
LDA-vector-SVM (RBF)	3.32	5.63	5.23	4.73

TABLE IV
minDCF RESULTS OF VARIOUS BASELINE SYSTEMS

	minDCF			
System	$G1$	$G2$	$G3$	AVE
i-vector	0.0564	0.0818	0.0790	0.0724
LDA-vector	0.0402	0.0601	0.0563	0.0522
i-vector-SVM (Linear)	0.0466	0.0699	0.0589	0.0585
i-vector-SVM (Poly)	0.0368	0.0681	0.0537	0.0529
i-vector-SVM (RBF)	0.0326	0.0601	0.0449	0.0459
LDA-vector-SVM (Linear)	0.0312	0.0545	0.0522	0.0460
LDA-vector-SVM (Poly)	0.0344	0.0559	0.0524	0.0476
LDA-vector-SVM (RBF)	0.0327	0.0556	0.0518	0.0467

1) IDR results: Note that in the OLR challenge, the target languages are known in prior, and the confidence scores are comparable across languages. This means that OLR can be treated as a language identification task, for which the language obtaining the highest score in a trail is regarded as the identification result. For such an identification task, IDR is a widely used metric [16], which treats errors on all languages equally serious. IDR is formally defined as follows:

$$IDR = \frac{T_c}{T_c + T_i}$$

where T_c and T_i are the numbers of correctly and incorrectly identified utterances, respectively. Table IV-E1 shows the IDR results of the baseline system. We can observe similar trends as

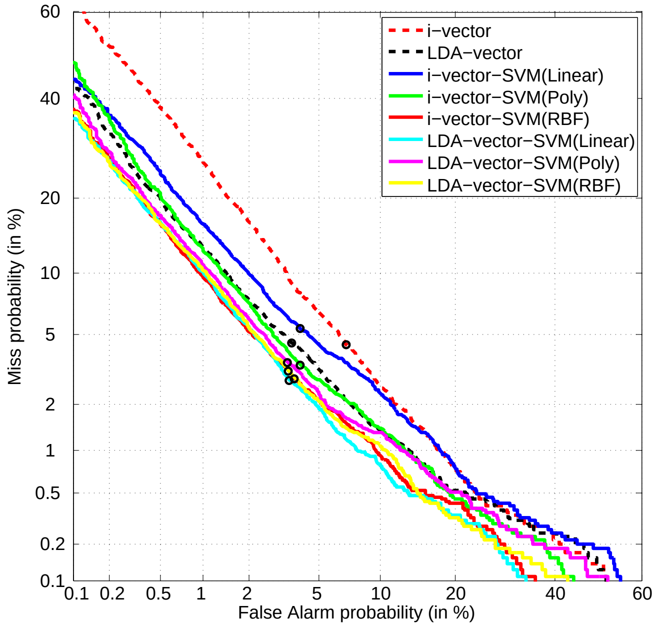

Fig. 3. The DET curves of various baseline systems with the data configuration $G1$.

with the verification metrics: C_{avg}, EER, minDCF and DET curve.

TABLE V
IDR RESULTS OF VARIOUS BASELINE SYSTEMS

	IDR%			
System	$G1$	$G2$	$G3$	AVE
i-vector	89.94	85.30	86.97	87.40
LDA-vector	91.37	87.31	88.41	89.03
i-vector-SVM (Linear)	89.18	83.96	86.60	86.58
i-vector-SVM (Poly)	90.78	83.62	87.16	87.19
i-vector-SVM (RBF)	92.19	85.91	89.59	89.23
LDA-vector-SVM (Linear)	92.11	87.76	88.03	89.30
LDA-vector-SVM (Poly)	91.66	87.50	87.76	88.97
LDA-vector-SVM (RBF)	91.82	87.51	87.79	89.04

V. Conclusions

We presented the data profile of the AP16-OL7 database that was released to support the AP16-OLR challenge on APSIPA 2016. The evaluation rules of the challenge was described, and a baseline system was presented. We show that the AP16-OL7 database is a suitable data resource for language recognition research.

References

[1] P. Sidwell and R. Blench, “14 the austroasiatic urheimat: the southeastern riverine hypothesis,” *Dynamics of human diversity*, p. 315, 2011.

[2] S. R. Ramsey, *The languages of China*. Princeton University Press, 1987.

[3] M. Shibatani, *The languages of Japan*. Cambridge University Press, 1990.

[4] B. Comrie, G. Stone, and M. Polinsky, *The Russian language in the twentieth century*. Oxford University Press, 1996.

[5] N. J. Enfield, “Areal linguistics and mainland southeast asia,” *Annual Review of Anthropology*, vol. 34, pp. 181–206, 2005.

[6] J. J. Godfrey, “Multilingual speech databases at ldc,” in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 23–26.

[7] T. Schultz, “Globalphone: a multilingual speech and text database developed at karlsruhe university.” in *INTERSPEECH*, 2002.

[8] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, “Speechdat-car. a large speech database for automotive environments.” in *LREC*, 2000.

[9] H. van den Heuvel, J. Boudy, Z. Bakcsi, J. Cernockỳ, V. Galunov, J. Kochanina, W. Majewski, P. Pollak, M. Rusko, J. Sadowski *et al.*, “Speechdat-e: five eastern european speech databases for voice-operated teleservices completed.” in *INTERSPEECH*, 2001, pp. 2059–2062.

[10] P. Roach, S. Arnfield, W. J. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel *et al.*, “Babel: an eastern european multi-language database.” in *ICSLP*, vol. 96, 1996, pp. 1892–1893.

[11] “The 2015 NIST language recognition evaluation plan (LRE09),” NIST, 2015, ver. 22-3.

[12] “The 2009 NIST language recognition evaluation plan (LRE09),” NIST, 2009, ver. 6.

[13] N. Dehak, P. G. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, “Language recognition via i-vectors and dimensionality reduction,” in *INTERSPEECH*, 2011, pp. 857–860.

[15] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Machine Learning Research*, 2008.

[16] B. Yin, E. Ambikairajah, and F. Chen, “Hierarchical language identification based on automatic language clustering.” in *INTERSPEECH*, 2007.